Rejoinder: Efficiency and Structure in MNIR

Matt Taddy, The University of Chicago Booth School of Business

I thank Prof. Blei and Grimmer for their comments; it is great to have one's work discussed by researchers who are both excellent statisticians and experts in their respective fields.

The discussion can be summarized under two themes. Prof. Blei is interested in extending MNIR to modeling additional, often latent, structure in text. Prof. Grimmer is concerned with causation and interpretability. Both will be answered in context of my original motivation for MNIR: the estimation efficiency derived from assumptions on $\mathbf{x}|y$. We'll begin with estimator properties in a simple illustration, then turn to discussion of latent factors and causal inference.

1 Efficiency

A related question of efficiency has been studied by Efron (1975) and Ng and Jordan (2002) in comparisons between logistic regression and 'generative' discriminant analysis. Efron's generative classifier applies Bayes rule to inverse multivariate normals $\mathbf{x}|y \sim \mathrm{N}(\mu_y, \Sigma)$, where $\mu_y = \mathbb{E}[\mathbf{x}|y]$ varies with $y \in \{0, 1\}$ but the covariance matrix is shared across populations. Given true normal covariate distributions separated by root Mahalanobis distances of 3 to 4, he finds predictions from this routine to be 1.5 to 3 times more efficient than logistic regression. This efficiency gain is smaller than that found by Ng and Jordan for a Naive Bayes algorithm (each covariate is fit as independent of the others given y), with their results loosely interpreted to imply $\log(n)$ times higher efficiency for the generative predictor. Although Naive Bayes independence is not assumed for the data itself, requirements on the amount of information about y available in each covariate have the effect of limiting conditional dependence.

Our model presents a third scenario: covariate dependence is fully specified via the negative correlation of a multinomial. Consider binary response $y \in \{0, 1\}$ and the joint word-sentiment distribution $\mathrm{p}(\mathbf{x}, y) = \mathrm{MN}(\mathbf{x} \mid \mathbf{q}(y))\mathrm{p}(y)$ where $q_j(y) = \exp[\alpha_j + \varphi_j y] / \sum_l \exp[\alpha_l + \varphi_l y]$ – that is, the collapsed model in Equation 1 of the main paper. Then the expected information for $\boldsymbol{\varphi}$ is $\pi\mathbf{W}$, where $\pi = \mathbb{E}[y]$ and $\mathbf{W} = \mathrm{diag}(\mathbf{q}_1) - \mathbf{q}_1\mathbf{q}_1'$ with $\mathbf{q}_1 = \mathbf{q}(y = 1)$, and standard results (e.g., van der Vaart, 1998, chap. 5) imply that in a fixed vocabulary the variance for maximum likelihood estimator $\hat{\boldsymbol{\varphi}}$ scales with $M = \sum_i \sum_j x_{ij}$, the total number of words.

PROPOSITION 1.1. *Assume the above joint model for y and $\mathbf{x}$ with $\pi > 0$, and write $\hat{\boldsymbol{\varphi}}$ for the MLE fit of $\boldsymbol{\varphi}$ in our collapsed MNIR model. The estimation error converges in distribution as*

$$\sqrt{\pi M}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \rightsquigarrow \mathrm{N}\left(\mathbf{0}, \mathbf{W}^{-1}\right)$$

Thus variance decreases with the amount of speech rather than with the number of speakers.

Prediction requires an accompanying forward model. If the collapsed model holds true, Bayes rule implies a forward predictor and results of Proposition 1.1 apply directly. A more realistic scenario has the collapsed model misspecified on an individual level. Consider a model of individual heterogeneity such that $\mathbf{x} \perp\!\!\!\perp y \mid \mathbf{x}'\boldsymbol{\varphi}, \mathbf{u}$ where $\boldsymbol{\varphi}$ can be estimated consistently as in Proposition 1.1 and $\mathbf{u}$ is a vector of unobserved random effects – for example, the model of Section 3.3 with $x_{ij} \sim \mathrm{Po}\left(\exp[\mu_j + \varphi_j y_i + u_{ij}]\right)$ and $y_i \perp\!\!\!\perp u_{ij} \sim \mathrm{N}(0,1)$. Write $z = \boldsymbol{\varphi}'\mathbf{f} = \boldsymbol{\varphi}'(\mathbf{x}/m - \frac{1}{n}\sum_i \mathbf{x}_i/m_i)$ for projection of mean shifted frequencies $\mathbf{F} = [\mathbf{f}_1 \cdots \mathbf{f}_n]'$, and say MNIR-OLS is the two-stage estimation of $\hat{\boldsymbol{\varphi}}$ in collapsed MNIR and $[\hat{\alpha}, \hat{\beta}]$ given $\hat{\mathbf{z}} = \mathbf{F}\hat{\boldsymbol{\varphi}}$ via least-squares (OLS). Consider the simple forward approximation $\mathbb{E}[y|\mathbf{f}, \mathbf{u}] = \alpha + \beta z$ (e.g., if $y = \tilde{\alpha} + \tilde{\beta} z + \boldsymbol{\gamma}'\mathbf{u} + \varepsilon$ and $u_j = a_j + b_j z + \nu_j$ with $\nu_j \perp\!\!\!\perp z$, then $\beta = \tilde{\beta} + \boldsymbol{\gamma}'\mathbf{b}$). Iterated expectation implies $\mathbb{E}\operatorname{argmin}_{\boldsymbol{\theta}} \sum_i (y_i - \alpha - \mathbf{f}_i'\boldsymbol{\theta})^2 = \mathbb{E}[y|\mathbf{f}] = \boldsymbol{\varphi}\beta$, such that OLS and MNIR-OLS have the same expectation and the effect of $\mathbf{u}$ on z is subsumed in β.

The distinction of MNIR-OLS is its estimation precision.

PROPOSITION 1.2. *Consider data from the joint word-sentiment distribution of Proposition 1.1 partitioned into documents $\{\mathbf{x}_i, y_i\}_{i=1}^n$ where $0 < \sum_i y_i < n$. Assuming a finite upper-bound for each $|\hat{\varphi}_j|$, the MNIR-OLS predictor $\hat{y}(\mathbf{x})$ for a new document $\mathbf{x}$ has*

$$\mathrm{var}\left(\hat{y}(\mathbf{x})\right) \xrightarrow{M\to\infty} \sigma^2\left(\frac{1}{n} + \frac{z^2}{\sum_{i=1}^n z_i^2}\right)$$

where $z = \mathbf{f}'\boldsymbol{\varphi}$ is the true projection for $\mathbf{x}$ and σ^2 is residual variance for regression of $\mathbf{y}$ on $\mathbf{z}$.

Proof. Note $\bar{z} = \mathbf{0}$ and $\mathrm{var}(\hat{y}(\mathbf{x})) = \mathrm{var}(\hat{\alpha}) + \mathbf{f}'\mathrm{var}(\hat{\boldsymbol{\varphi}}\hat{\beta}_{\hat{\mathbf{z}}})\mathbf{f}$ where $\hat{\beta}_{\hat{\mathbf{z}}}$ is OLS slope on $\hat{\mathbf{z}} = \mathbf{F}\hat{\boldsymbol{\varphi}}$. From Proposition 1.1 and the continuous mapping theorem we have $\hat{\boldsymbol{\varphi}} \xrightarrow{\mathrm{p}} \boldsymbol{\varphi}$ and $\hat{\beta}_{\hat{\mathbf{z}}} \rightsquigarrow \hat{\beta}_{\mathbf{z}}$. Slutsky's lemma yields $\hat{\boldsymbol{\varphi}}\hat{\beta}_{\hat{\mathbf{z}}} \rightsquigarrow \boldsymbol{\varphi}\hat{\beta}_{\mathbf{z}}$ with variance $\boldsymbol{\varphi}\mathrm{var}(\hat{\beta}_{\mathbf{z}})\boldsymbol{\varphi}' = \sigma^2\boldsymbol{\varphi}\boldsymbol{\varphi}'/\sum_i z_i^2$. Given that $\hat{\boldsymbol{\varphi}} \mapsto \hat{\boldsymbol{\varphi}}\hat{\beta}_{\hat{\mathbf{z}}}$ is bounded on its finite domain, the Portmanteau lemma implies our convergence.

Thus, in our simple cartoon, MNIR-OLS approaches with *number-of-words* the error rate of univariate least-squares. This holds for infill (where n is constant but speech-per-document grows) as well as when n is growing with M and the right-hand-side of 1.2 is decreasing. Regularized estimation, say as applied in the main article, should help efficiency in tougher

setups (e.g., where vocabulary grows with M) but will increase bias. Although we've focused on linear models many other options are available – for example, tree methods (e.g., Breiman, 2001) work well in low dimensions for nonlinearity and variable interaction. The principles remain the same: results like Proposition (1.1) show efficiency in collapsed IR, and one hopes to be able to account for individual-level misspecification in the low dimensional forward model.

2 Latent factors

Prof. Blei's 2nd extension is an especially promising idea. Random effects were originally viewed as a nuisance necessary for understanding misspecification. However, a low-dimensional latent factorization of these effects would be a powerful tool for exploration and prediction. It provides a middle ground between LDA and MNIR.

Such a model has log-odds $\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Phi}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{u}$ where $\mathbf{u} = [u_1 \dots u_K]'$ is a K-dimensional factor vector. $\boldsymbol{\Gamma}$ can then be interpreted as logit-transformed LDA topics for variation in text not explained by variables in $\mathbf{y}$. Just as $\boldsymbol{\Phi}'\mathbf{x}$ is sufficient for $\mathbf{y}$, the topic projection $\boldsymbol{\Gamma}'\mathbf{x}$ will be sufficient for latent factors. Therefore the model provides both a new way to think about latent structure in text and a strategy for fast computation of topic weights.

The difficulty with latent factor modeling is estimation. On the one hand, although the model is more complex, estimation variance should still decrease with M because of the multinomial assumption on $\mathbf{x}$ (indeed, similar arguments can explain the solid performance of LDA and sLDA regression). However, there are two big computational issues in posterior maximization with document-specific $\boldsymbol{\Gamma}\mathbf{u}_i$: you can no longer collapse the likelihood, and you need to jointly solve for $\boldsymbol{\Gamma}$ and $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_n]'$. Since the discussants and I work on corpora many orders larger than the examples in this article, additional latent structure is only useful if we can devise scalable algorithms for its estimation.

On the lack of collapsibility, which is also an issue for high-dimensional $\mathbf{y}$, I have had success applying a MapReduce strategy (Dean and Ghemawat, 2004). A factorized likelihood is obtained by assuming counts x_{ij} and x_{ik} for $j \neq k$ are independent and Poisson distributed given $\mathbf{y}_i$ and $\mathbf{u}_i$ (centered on intensity $\exp(m_i/p)$ for convenience). The Map step groups counts on each column of $\mathbf{X}$ (i.e., for each word) and the Reduce step is a (possibly zero-inflated) Poisson log regression of each word count onto $\mathbf{y}_i$ and $\mathbf{u}_i$. Exponential family parametrization of the Poisson allows the same sufficiency results, and the multinomial distribution for vectors of independent Poissons given their sum implies a close connection to MNIR. A paper on this approach to distributed multinomial regression is under preparation.

Even with these parallel algorithms, it is difficult to solve for both $\mathbf{U}$ and $\mathbf{\Gamma}$. A fixed-point solver (iterating between maximization for each conditional on the other) is usually too slow. One could impute a rough guess for $\mathbf{U}$ (e.g., from a PCA of document tf-idf), but this is only a stand-in solution. Recent advances in distributed optimization using ADMM (Boyd et al., 2010) may offer a way forward, iterating from unique $\mathbf{U}_j$ for each j^{th} word towards shared $\mathbf{U}$ across vocabulary, but this is just conjecture. The problem of latent factor MNIR for large corpora remains unsolved. I look forward to further discussion with Prof. Blei on this because it is something that his lab, if anybody, has a good chance of tackling.

3 Interpretability

Prof. Grimmer's comments are focused on interpretability: the translation from estimated models to scientific mechanisms. In particular, he and other social scientists are interested in questions of *causation*. This is among the toughest of topics in statistics, and one that is only growing in both difficulty and importance with the amount and dimension of our data.

First, we should not underestimate the importance of predictive ability in causal modeling. The goal is always good prediction, but to understand causation we want a model that predicts well when one covariate changes and all others stay constant. Some of the best causal inference schemes are explicitly predictive: matching, treatment-effects models, and propensity scores rely upon estimation of the rate at which treated individuals were assigned to that group. As an example, colleagues and I are interested in measuring attribution for digital advertisements (i.e., how an ad *causes* changes in consumer behavior). This is a notoriously tough problem, since the fact that a consumer sees an ad is highly correlated with the likelihood that they were already looking to buy a certain product. MNIR for a consumer's text (e.g. on social media) and their browser history (where website counts are treated like word counts) can be used to efficiently predict the probabilities both that they see an ad and that they buy a product, and we hope to use this to disentangle these correlated outcomes.

However, instead of using text to help control for unobserved variables, Prof. Grimmer is seeking methods to infer the mechanisms behind word choice. This is because he rightly wants to ensure that word loadings correspond to a general notion of partisanship – one that is portable between, say, newspapers and congressional speech. This is the causal problem exploded to simultaneous inference for thousands of correlated outputs. Regardless, MNIR is a natural starting point: I assume that 'sentiment' causes speech rather than the inverse. From this one can look to apply the structural models used in econometrics and biostatistics. As mentioned,

the effects of other inputs are 'controlled for' by including them in the log-odds, say as $\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\varphi} y + \boldsymbol{\Theta}\mathbf{v}$ where $\mathbf{v} = [v_1 \dots v_d]'$ are confounding variables. Going further, an MNIR treatment effects estimator would regress y on $\mathbf{v}$ and include the fitted expectation in the equation for $\boldsymbol{\eta}$. One needs to be careful here, as techniques used for efficiency in high dimensions, such as sparse regularization, can bias inference in unexpected ways. See Belloni et al. (2012) for recent work on sparse high-dimensional treatment effects estimation.

Finally, we should be aware of the limits of frameworks like MNIR (this also relates to Prof. Blei's 3rd extension). As Prof. Grimmer says, it *is* difficult to know what covariates should be included or excluded from the model. However, this will always be as much of a problem in text analysis as it has long been in social science. The 'what' that we measure is only ever defined in terms of observables and the model assumed around them (even with human coders sentiment is dictated by the questions we ask). The goal is to have this be as close as possible to our abstract ideal. For example, an ongoing project at Booth is investigating the history of partisanship in congressional speech. To define partisanship, we look at average predictability of party identity given words drawn from the distribution of speech for a given party. The question of partisanship has been transformed to one of predictability, and this notion is refined by controlling for causes of word choice (e.g., geography, race) that we understand as distinct from partisanship. It is healthy to keep this inference separate from abstract meanings for sentiment or partisanship, in order to be clear on where evidence ends and speculation begins.

Thanks to Jesse Shapiro, Matt Gentzkow, and Christian Hansen for helpful discussion.

References

Belloni, A., V. Chernozhukov, and C. Hansen (2012). Inference on treatment effects after selection amongst high-dimensional controls. MIT Department of Economics Working Paper No. 12-13.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*, 1–122.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Dean, J. and S. Ghemawat (2004). MapReduce: Simplified data processing on large clusters. In *Proceedings of Operating Systems Design and Implementation*, pp. 137–150.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* (70), 892–898.

Ng, A. Y. and M. I. Jordan (2002). On discriminative vs generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems (NIPS)*.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge.